EMPLOYEE ATTRITION:
A STUDY ON FEATURES OF SIGNIFICANCE & PREDICTION MODEL BUILDING

By

Md. Sharafat Hossain

ID: 183-44-172

This report presented in fulfillment of the requirements for the Degree of M.Sc. in Software Engineering

Department of Software Engineering
Daffodil International University

SPRING 2019

LETTER OF APPROVAL

This is a study on various employee attrition factors to see how they can be used to predict employee attrition for an organization, based on an HR dataset available online for open use. An analysis using multiple feature selection methodology to select the most important features that contribute in employee leaving the organization, and thereby conclude with a prediction model is submitted by Md. Sharafat Hossain to the Department of Software Engineering, Daffodil International University, has been accepted as fulfillment of the requirements for the degree of M.Sc. and approved as to style contents.

BOARD OF EXAMINERS

Dr. Touhid Bhuiyan Head
Professor & Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Examiner Internal

Examiner Internal

Examiner External

ACKNOWLEDGEMENT

I am grateful to my creator for giving me the opportunity to complete this research work and learn so much. I am thankful to my research supervisor, Mr. K. M. Imtiaz-Ud-Din, for providing careful guidance starting from selecting the research scope to successfully finalizing the research work. I would also like to thank Mr. M. Khaled Sohel for his valuable comments which was always insightful. Finally, I want to express my gratitude to Professor Dr. Touhid Bhuiyan, head of the Software Engineering faculty, for inspiring us in all means.

THESIS DECLARATION

I hereby declare that, this thesis report is done by me under the supervision of Mr. K. M. Imtiaz-Ud-Din, Assistant Professor, Department of Software Engineering, Daffodil International University, in fulfillment of my original work. I am also declaring that according to the best of my knowledge, neither this thesis nor any part therefore has been submitted else here for the award of M.Sc. or any degree.

Signature of the Project Supervisor

02/05/19

Signature of the Internship Supervisor

Supervised by

K. M. Imtiaz-Ud-Din
Assistant Professor
Department of Software Engineering
Daffodil International University

Submitted by

Md. Sharafat Hossain
ID: 183-44-172
Department of Software Engineering
Daffodil International University

ABSTRACT

Employee attrition is a significant problem for any organization. In true sense, the cost of replacing a well-trained or an employee with good performance can be really big, sometimes even more than what makes sense about money. Along with the time spent for candidate interview, selection, notice period, joining bonus etc. the loss of production hours for a significant amount of time while the new employee gets used to the system and generates output combines in a vital amount of loss for the organization. Employee turnover happening on a regular basis causes the organization to decrease in its collective knowledge base and experience gathered over time. Also, possibilities of errors and issues increases at a great rate with new workers until they gather enough knowledge on the system.

Being able to understand what factors contribute most for an employee leaving can help the management to plan according actions to improve employee retention possibility as well as to plan a new hire in advance. Our research is aimed to discover that can data science help find out which attributes from the dataset contribute more in an employee leaving his/her organization. Also, we will try to build a prediction model which will predict whether an employee will be leaving the organization so that the management can take appropriate steps to reduce employee turnover. In this research we selected some of the most widely used feature selection techniques, and applied them on a selected dataset. Then we applied a clustering algorithm on the dataset to identify how the selected features affect employee attrition. Finally, we have built a prediction model for employee attrition and measured its performance. We have discussed the results to know how the present research expands on previous works done on this topic and based on our limitations – recommendations on how future work can move forward.

TABLE OF CONTENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Explanation

CHAPTER 1

INTRODUCTION

1.1 Background

Employee attrition is a significant problem for any organization. In true sense, the cost of replacing a well-trained or an employee with good performance can be really big, sometimes even more than what makes sense about money. Along with the time spent for candidate interview, selection, notice period, joining bonus etc. the loss of production hours for a significant amount of time while the new employee gets used to the system and generates output combines in a vital amount of loss for the organization. What if management of the organization could scientifically anticipate which employees are leaving and the reason behind their turnover?

To understand why employee leaves any organization with the help of data science, first we need to focus on which factors contribute most in the employee's decision. Feature selection as well as data cleaning must be the considered as the first and most important step of any model designing. Feature selection is the method of selecting the features that contribute most to the prediction variable or output in which the model is concerned with. Inclusion of irrelevant features in the data can highly decrease the accuracy of the model. It is a general conception that employee attrition is generally affected by attributes like pay, tenure, age, job satisfaction, working conditions, growth potential etc.

Correlation between the factors can also play a significant role in analyzing how the factors behave with each other inside the model. Correlation is a statistical measure that provides the information how one variable is interrelated to other variables.

1.1.1 Feature Selection Methods

Univariate Selection:

In Univariate Selection, statistical tests are be applied for selecting the features which have the strongest association with the output variable. To implement univariate selection, the scikit-learn library offers the **SelectKBest** class. It is used with a collection of various statistical tests to select a specific list of features.

Feature Importance:

The feature importance of each single feature of the dataset can be obtained by using the feature importance property of the model. Feature importance provides a score for each feature of the data. A higher score means the feature is more important or relevant to the output variable. Feature importance is an inbuilt class, coming with Tree Based Classifiers.

1.1.2 Correlation

Correlation is one of the well-known statistical tools that provides the information about how the variables inside the model are interrelated to each other. In other words, it is a measure of the degree to which changes in the value of one variable is used to predict changes in the value of another variable.

Correlation Coefficient:

The correlation coefficient is a statistical measurement tool that is used to calculate the relationship strength between the relative activities of two variables. The values always range between **-1.0** and **1.0**. A calculated number is found to be greater than 1.0 or less than -1.0, it means that an error must have happened in the correlation measurement calculation. A correlation of -1.0 portrays a perfect negative correlation, while a correlation of 1.0 demonstrations a perfect positive correlation.

A correlation of 0.0 means that no relationship exists between the movements of the two variables. Below is a sample interpretation of correlation coefficient:

Coefficient Value	Interpretation
-1	A perfect negative relationship
-0.7	A strong negative relationship
-0.5	A moderate negative relationship
-0.3	A weak negative relationship
0	No linear relationship
0.3	A weak positive relationship
0.5	A moderate positive relationship
0.7	A strong positive relationship
1	A perfect positive relationship

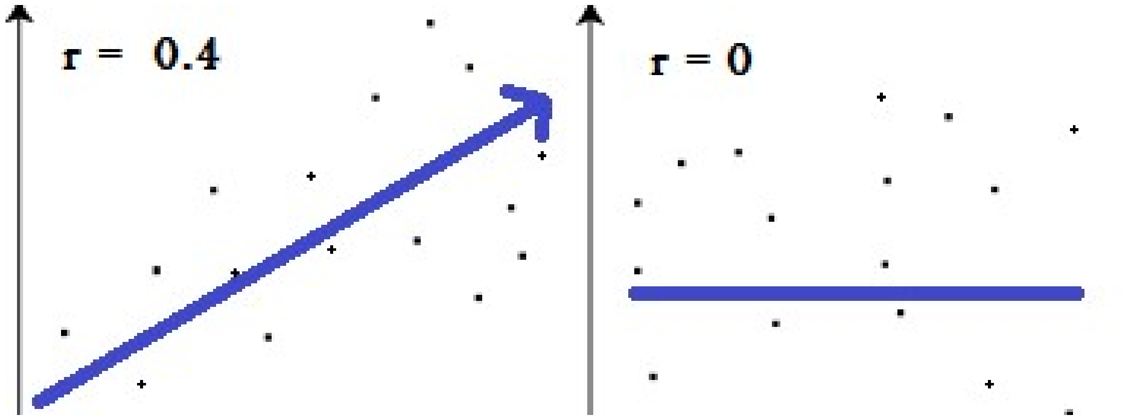


Figure01: Correlation Coefficient

Calculation:

One of the most widely used correlation coefficient, Pearson product-moment correlation is calculated in below steps:

- The covariance of the two variables in question is determined.
- Each variable's standard deviation is calculated.
- The correlation coefficient is calculated by dividing the covariance by the product of the two variables' standard deviations.

Advantages:

- Can show strength of relationship between two variables
- Study behavior that is generally can't be studied
- Gain quantitative data which can be easily analyzed

Predictive Modeling:

Predictive modeling is a process that makes use of mathematical methods and probability to make prediction of an event or outcome. Each model is created with of a number of predictors, which can be described as variables that are expected to impact the future output. A mathematical approach makes use of an equation-based model that describing the phenomenon under attention. The model is used to predict a result at a future state or time depending on changes to the model's inputs variables. The model parameters help describe how the model inputs are going to influence the outcome.

Gradient Boosting Classifier:

Gradient boosting is known as one of the most accepted and powerful techniques used for predictive model building. Boosting refers to renovating weak learners into strong learners.

The general idea is to training a decision tree so that each new tree is a fit on a modified version of the original data set. This process is repeated for a specified number of iterations. Subsequent trees help us to categorize observations that are not well classified by the previous trees. Forecasts of the final collaborative model is therefore the weighted sum of the predictions made by the previous tree models.

Gradient boosting involves three elements:

- A loss function to be optimized.
- A weak learner to make predictions.
- An additive model to add weak learners to minimize the loss function.

Gradient boosting identifies the shortcomings by using gradients in the loss function (*y=ax+b+e , e needs a special mention here since it is the error term)*. The loss function is described as a measure to specify how well the model's coefficients are fitting the original dataset.

Advantages:

- Often provides predictive accuracy that cannot be beat.
- Lots of flexibility - can optimize on different loss functions and provides several hyperparameter tuning options that make the function fit very flexible.
- No data pre-processing required - often works great with categorical and numerical values as is.
- Handles missing data - imputation not required.

1.2 Motivation of The Research

Employee attrition is a significant problem for any organization. In true sense, the cost of replacing a well-trained or an employee with good performance can be really big, sometimes even more than what makes sense about money. Along with the time spent for candidate interview, selection, notice period, joining bonus etc. the loss of production hours for a significant amount of time while the new employee gets used to the system and generates output combines in a vital amount of loss for the organization. Employee turnover happening on a regular basis causes the organization to decrease in its collective knowledge base and experience gathered over time. Also, possibilities of errors and issues increases at a great rate with new workers until they gather enough knowledge on the system.

Being able to understand what factors contribute most for an employee leaving can help the management to plan according actions to improve employee retention possibility as well as to plan a new hire in advance. Coming from the corporate industry myself, our research wanted to discover that whether data science is able to help find out which attributes from the dataset contribute more in an employee leaving the organization and help the management.

1.3 Problem Statement

A research problem is an area of concern in the existing knowledge that points to the need for further understanding and investigation. As seen from the literature review, the previous researchers have not collectively paid enough concentration on collaborating and finding contributing factors for employee attrition and use them in prediction analysis.

1.4 Research Questions

The research questions for our research are:

RQ1: Is it possible to predict how much likely an active employee will be leaving the organization?

RQ2: What are the significant features/attributes of an employee quitting the organization?

1.5 Research Objective

The research objective is to find answers to our research questions. Therefore, our research objectives are:

- To find out if data science/predictive analysis techniques can contribute in an employee retention/attrition model.
- Evaluate and find out which features or attributes are most important for predicting an employee attrition.

1.6 Research Scope

The scope of our research was limited to our dataset for HR analytics, and the employee attributes available in the dataset. There may be other existing factors that contribute to employee attrition. However, we limited our research on the used dataset only.

1.7 Thesis Organization

In the following sections, we first discussed about previous literature including the research gap. Secondly, we addressed research methodology and our research model. Then, we provided research results and discussions. Finally, we concluded by discussing on recommendations with findings, limitations and future directions.

CHAPTER 2

LITERATURE REVIEW

2.1 Previous Literature

According to **Boswell, Boudreau and Tichy (2005)**, the decision of leaving the Organization is not easy for an individual employee as well as significant energy is spent on finding new jobs, adjusting to new situations, giving up known routines and interpersonal connection and is so stressful. Therefore, if timely and proper measures are taken by the Organizations, some of the voluntary turnover in the Organization can be prevented.

Trevor, (2001), in his research found that employees who perform better and are intelligent enough have more external employment opportunities available compared to average or poor performance employees and thus they are more likely to leave. High rates of voluntary turnover of such employees are often found to be harmful or disruptive to firm's performance **(Glebbeck & Bax, 2004)**. When poor performers, choose to leave the Organization, it is good for the Organization **(Abelson & Baysinger, 1984)**. Further voluntary turnover of critical work force is to be differentiated into avoidable and unavoidable turnover **(Barrick & Zimmerman, 2005)**.

Goodwill of the company gets hampered due to more employee turnover rate and the competitors start poking their nose to recruit best talents from them. Efficiency of work is hampered to a large extent. **(Dhobal & Nigam, 2018)**

http://www.iosrjournals.org/iosr-jbm/papers/Vol20-issue2/Version-4/A2002040127.pdf

Feature Selection is one of the core concepts in machine learning which hugely impacts the performance of your model. The data features that you use to train your machine learning models have a huge influence on the performance you can achieve. Irrelevant or partially relevant features can negatively impact model performance. **(Sheikh, 2018)**

https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e

Feature selection is a process where you automatically select those features in your data that contribute most to the prediction variable or output in which you are interested. Having irrelevant features in your data can decrease the accuracy of many models, especially linear algorithms like linear and logistic regression. **(Brownlee, 2016)**

https://machinelearningmastery.com/feature-selection-machine-learning-python/

Correlation is defined as *a relation existing between phenomena or things or between mathematical or statistical variables which tend to vary, be associated, or occur together in a way not expected by chance alone* by the Merriam-Webster dictionary. The relationship (or the correlation) between the two variables is denoted by the letter r and quantified with a number, which varies between −1 and +1. Zero means there is no correlation, where 1 means a complete or perfect correlation. The sign of the r shows the direction of the correlation. A negative r means that the variables are inversely related. The strength of the correlation increases both from 0 to +1, and 0 to −1. **(Emerg Med, 2018)**

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6107969/

Gradient boosting machines are a family of powerful machine-learning techniques that have shown considerable success in a wide range of practical applications. They are highly customizable to the particular needs of the application, like being learned with respect to different loss functions. (Natekin & Knoll, 2013)

https://www.researchgate.net/publication/259653472_Gradient_Boosting_Machines_A_Tutorial/download

2.2 Research gap

- A review of previous literature advises that not too many research works have been carried out to study the combination of feature selection and prediction model building for employee attrition.

2.3 Summary

Our study has been motivated by the need for a research on how data science can help an organizations management in the area of employee retention or attrition. We tried to demonstrate which factors contribute more on an employee leaving the organization and if that can be forecasted so that management can take necessary steps to reduce loss earlier.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Feature Selection Methods

For our research, we decided to see which employee attributes are found to be contributing most for their attrition from the organization. We selected Univariate Selection and Feature Importance as our feature selection methods. From there, we analyzed which outcomes from both selection methods are common. This means, these outcomes are having the most influence an employee leaving the organization from our chosen dataset.

3.2 Correlation

After the selection of the most important features from the dataset was done, we calculated the correlation coefficients between those features. This allowed us to study how the features are interrelated and how they affect the value of the final prediction variable (positively or negatively).

3.3 Gradient Boosting Classifier

A prediction model on our selected dataset is build using Gradient Boosting Method. We used scikit-learns **train_test_split ()** method to get a training dataset and a testing dataset for our model. The training dataset (60% of total data) is used to train the model and the testing dataset (40% of total data) is used to test the model.

The performance of the model is measured by using scikit-learns evaluation metrics accuracy, precision and recall.

3.4 Data Source

The data used for this research is available in Kaggle.

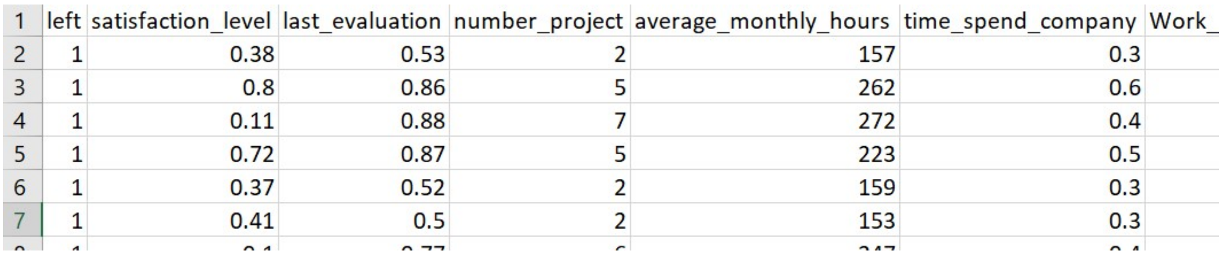
https://www.kaggle.com/gummulasrikanth/hr-employee-retention

3.5 Data Description:

For our research work, the data we had considered, consisted of various employee attributes while working at the company.

In the used dataset, each row represents an employee, each column contains employee attributes:

- satisfaction_level (0–1)
- last_evaluation (years since last evaluation was done)
- number_projects (Number of completed projects during work)
- average_monthly_hours (average time spent at workplace)
- time_spend_company (years spent at the company)
- Work_accident (if the employee had any accident at the workplace)
- left (if the employee left the workplace (1/ 0))
- promotion_last_5years (if the employee was promoted within the previous five years)
- Department (apartment which employees work in)
- salary (salary at relative level – low/medium/high)



1	left	satisfaction_level	last_evaluation	number_project	average_monthly_hours	time_spend_company	Work_
2	1	0.38	0.53	2	157	0.3	
3	1	0.8	0.86	5	262	0.6	
4	1	0.11	0.88	7	272	0.4	
5	1	0.72	0.87	5	223	0.5	
6	1	0.37	0.52	2	159	0.3	
7	1	0.41	0.5	2	153	0.3	

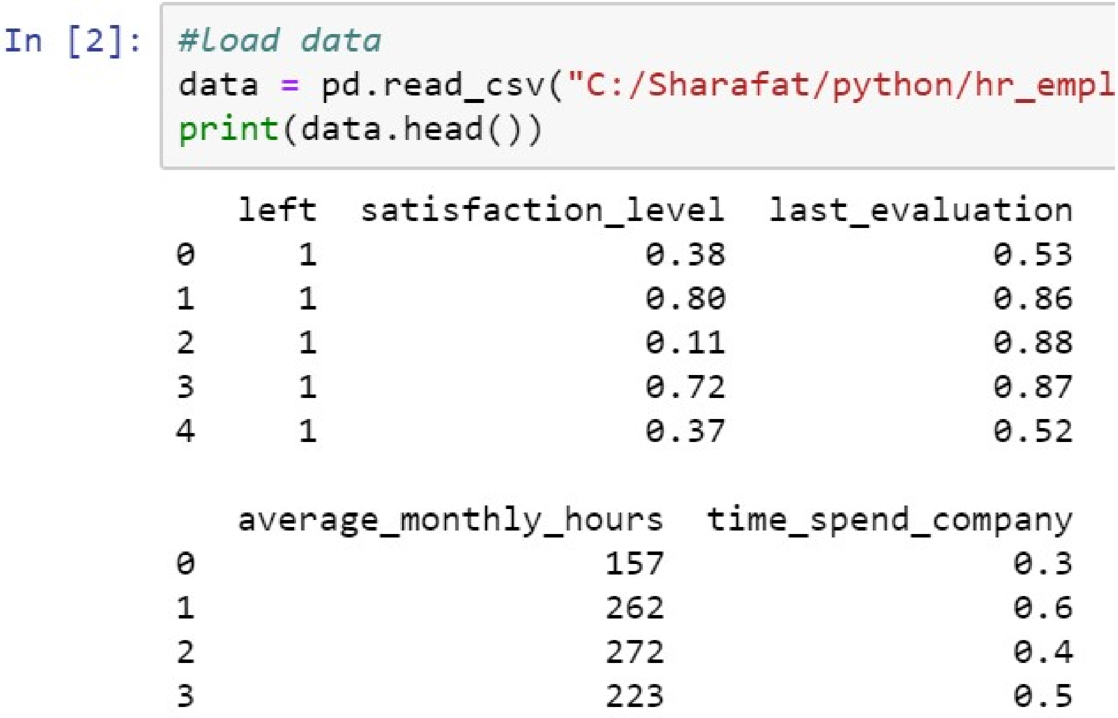
Figure02: Raw Data Preview

3.6 Data Preprocessing

Before fitting in the model, the data had to be preprocessed and cleaned. The data loading and preprocessing steps are described below.

Data Loading:

The selected HR dataset was loaded using pandas's **read_csv ()** function.

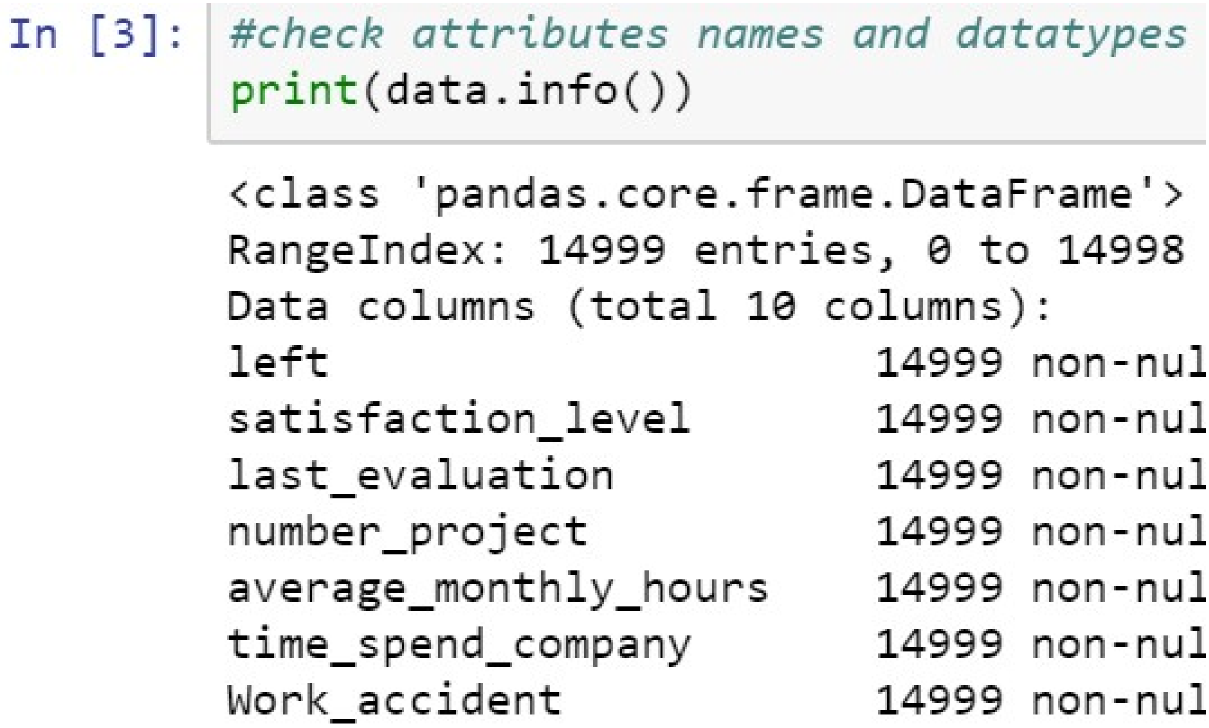


```
In [2]: #Load data
        data = pd.read_csv("C:/Sharafat/python/hr_empl
        print(data.head())
```

```
   left  satisfaction_level  last_evaluation
0     1                0.38             0.53
1     1                0.80             0.86
2     1                0.11             0.88
3     1                0.72             0.87
4     1                0.37             0.52

   average_monthly_hours  time_spend_company
0                    157                 0.3
1                    262                 0.6
2                    272                 0.4
3                    223                 0.5
```

Figure03: Data Load Preview

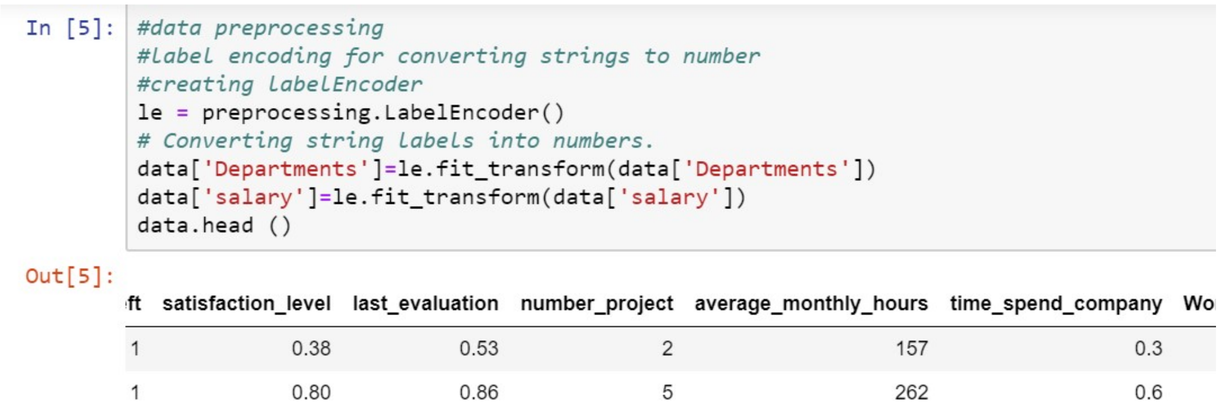
Then we took a look into the data attribute names and datatypes using **info()**



```
In [3]: #check attributes names and datatypes
        print(data.info())

        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 14999 entries, 0 to 14998
        Data columns (total 10 columns):
        left                         14999 non-nul
        satisfaction_level           14999 non-nul
        last_evaluation              14999 non-nul
        number_project               14999 non-nul
        average_monthly_hours        14999 non-nul
        time_spend_company           14999 non-nul
        Work_accident                14999 non-nul
```

Figure04: Data Attributes and Datatypes Preview

Generally, machine learning algorithms prefers numerical input data. Therefore, we needed to convert the categorical columns into numerical columns. We mapped each value to a number. For example, the Salary column's value has been represented by **low:0, medium:1, and high:2**. This procedure is called label encoding. We used sklearns **LabelEncoder ()** to get this done.



In [5]:

```
#data preprocessing
#label encoding for converting strings to number
#creating labelEncoder
le = preprocessing.LabelEncoder()
# Converting string labels into numbers.
data['Departments']=le.fit_transform(data['Departments'])
data['salary']=le.fit_transform(data['salary'])
data.head ()
```

Out[5]:

ft	satisfaction_level	last_evaluation	number_project	average_monthly_hours	time_spend_company	Wo
1	0.38	0.53	2	157	0.3	
1	0.80	0.86	5	262	0.6	

Figure05: Label Encoding

3.7 Summary

Univariate Selection and Feature Importance have been selected as the feature selection methods for employee attrition from the used HR dataset. Correlation coefficient has been calculated with on data by the features that will be common from both feature selection techniques. Finally, a prediction model based on the dataset is built and its performance measured based on accuracy, precision and recall.

CHAPTER 4

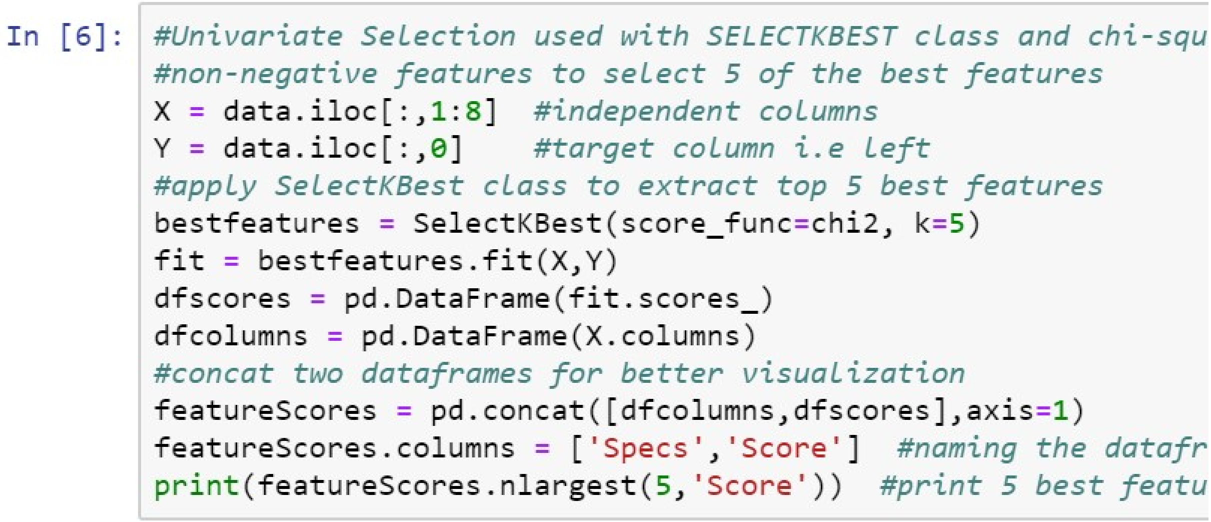
RESULTS AND DISCUSSIONS

4.1 Data Analysis Techniques

4.1.1 Univariate Selection:

In Univariate Selection, statistical tests are be applied for selecting the features which have the strongest association with the output variable. To implement univariate selection, the scikit-learn library offers the **SelectKBest** class. It is used with a collection of various statistical tests to select a specific list of features.

For our case, we used the chi squared (chi^2) statistical test for non-negative features to select 5 of the best features from our dataset. Here 5 best features mean the columns in the dataset that has most influence in determining the values of the column '**left**' which indicates whether the employee has left the organization (1) nor not (0).



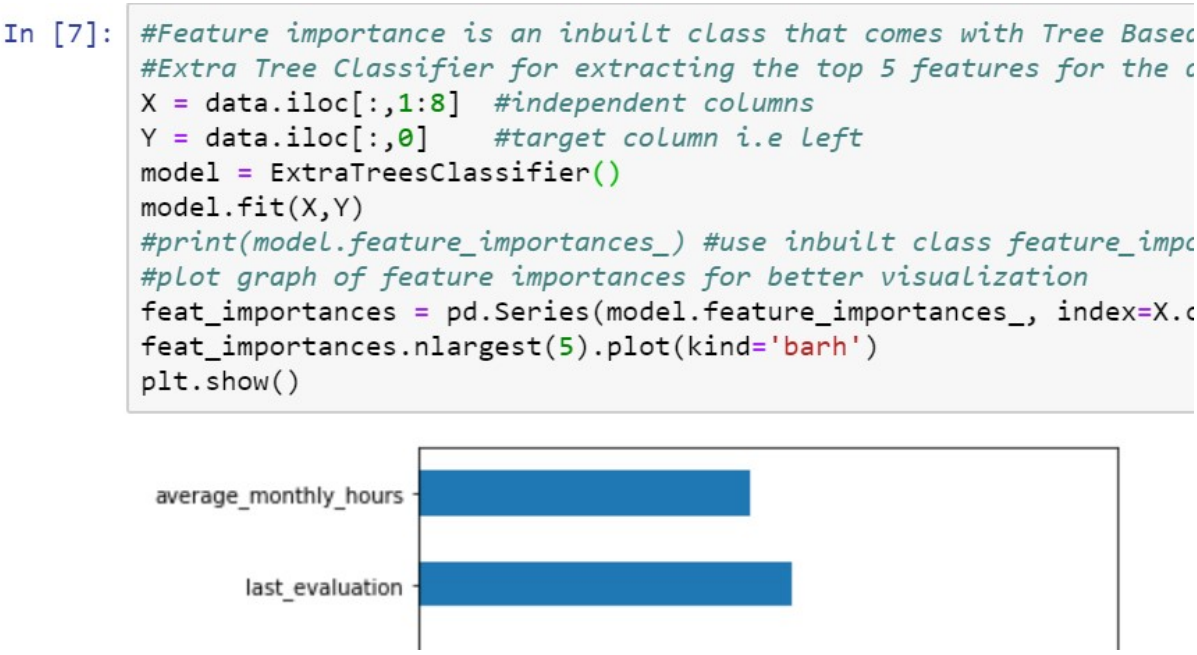
```
In [6]: #Univariate Selection used with SELECTKBEST class and chi-squ
        #non-negative features to select 5 of the best features
        X = data.iloc[:,1:8]  #independent columns
        Y = data.iloc[:,0]    #target column i.e left
        #apply SelectKBest class to extract top 5 best features
        bestfeatures = SelectKBest(score_func=chi2, k=5)
        fit = bestfeatures.fit(X,Y)
        dfscores = pd.DataFrame(fit.scores_)
        dfcolumns = pd.DataFrame(X.columns)
        #concat two dataframes for better visualization
        featureScores = pd.concat([dfcolumns,dfscores],axis=1)
        featureScores.columns = ['Specs','Score']  #naming the datafr
        print(featureScores.nlargest(5,'Score'))  #print 5 best featu
```

Figure06: Univariate Selection

4.1.2 Feature Importance:

The feature importance of each single feature of the dataset can be obtained by using the feature importance property of the model. Feature importance provides a score for each feature of the data. A higher score means the feature is more important or relevant to the output variable. Feature importance is an inbuilt class, coming with Tree Based Classifiers.

For our case, we constructed an ExtraTreesClassifier classifier on our dataset to identify the columns in the dataset that has most influence in determining the values of the column '**left**' which indicates whether the employee has left the organization (1) nor not (0).



```
In [7]: #Feature importance is an inbuilt class that comes with Tree Basec
        #Extra Tree Classifier for extracting the top 5 features for the c
        X = data.iloc[:,1:8]  #independent columns
        Y = data.iloc[:,0]    #target column i.e left
        model = ExtraTreesClassifier()
        model.fit(X,Y)
        #print(model.feature_importances_) #use inbuilt class feature_impc
        #plot graph of feature importances for better visualization
        feat_importances = pd.Series(model.feature_importances_, index=X.c
        feat_importances.nlargest(5).plot(kind='barh')
        plt.show()
```

Figure07: Feature Importance

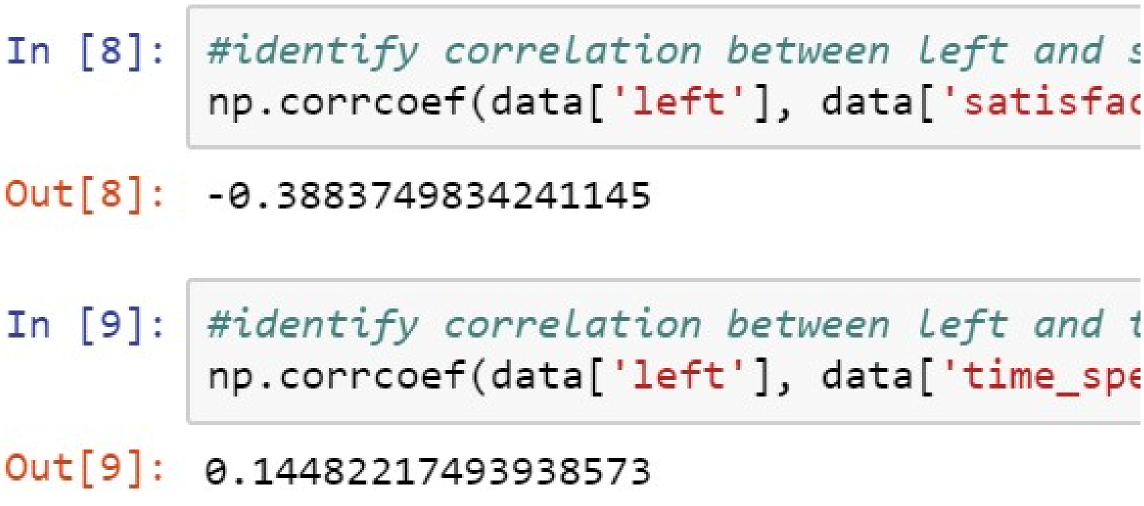
4.1.3 Correlation

In statistical terms, correlation is used to denote association between two quantitative variables. It is also assumed that the association is linear, that one variable increase or decrease a fixed amount for a unit increase or decrease in the other.

Now we calculated the correlation coefficients between the predictor variable of our dataset (employees who left the organization) denoted by the column ‘**left**’ with the common set of features (from two feature selection methods) to have the most influence on the value of ‘**left**’.

The common features are found to be:

- average_monthly_hours
- satisfaction_level
- time_spend_company

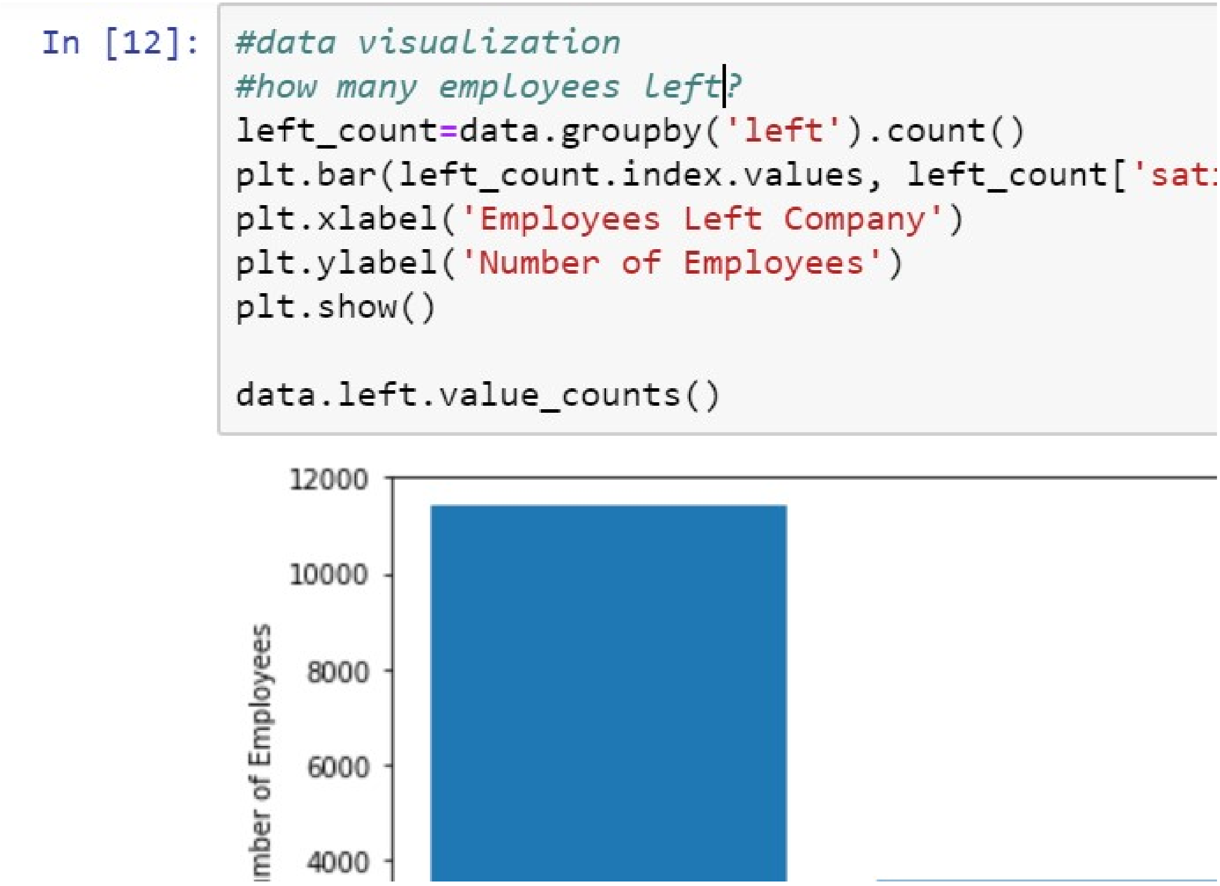


```
In [8]:  #identify correlation between left and s
         np.corrcoef(data['left'], data['satisfac

Out[8]:  -0.3883749834241145

In [9]:  #identify correlation between left and t
         np.corrcoef(data['left'], data['time_spe

Out[9]:  0.1448217493938573
```

Figure08: Correlation Coefficients

4.2 Data Analysis

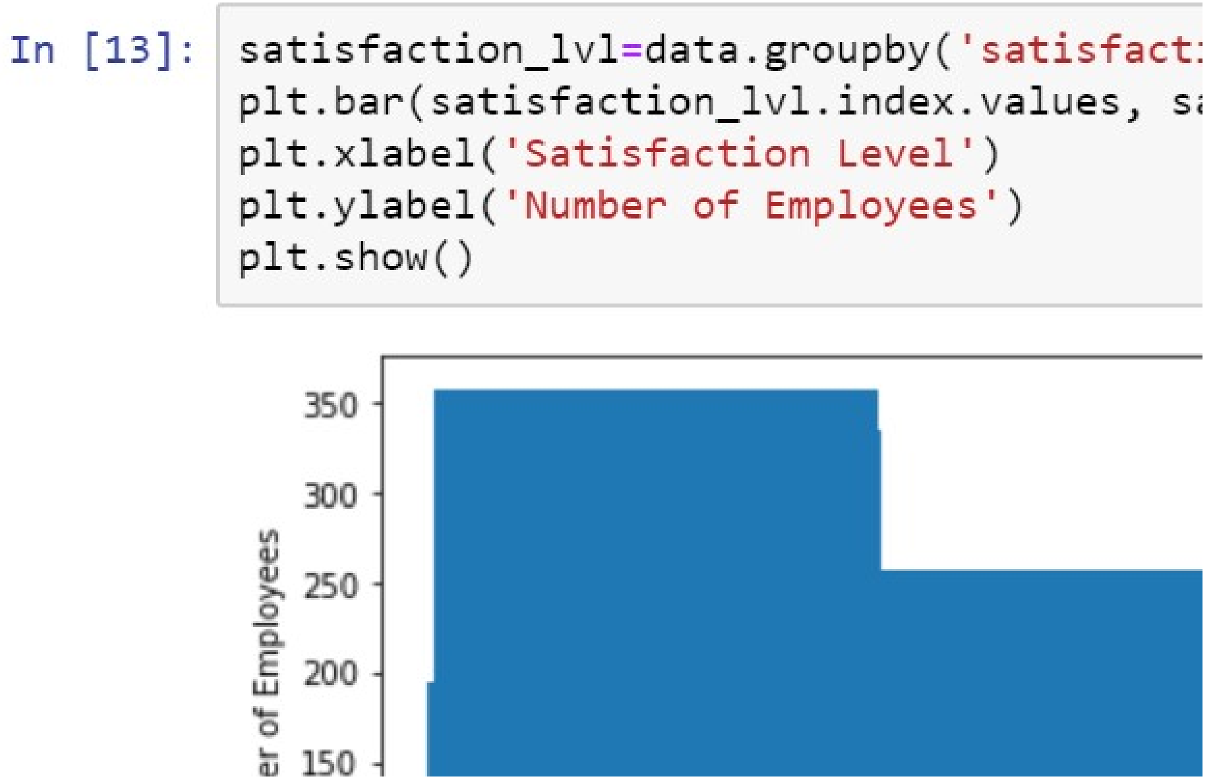
To see how many employees from the dataset actually left the organization, we calculated and found that the no of employee left is 23 % of the total employment.



```
In [12]: #data visualization
         #how many employees left?
         left_count=data.groupby('left').count()
         plt.bar(left_count.index.values, left_count['sat
         plt.xlabel('Employees Left Company')
         plt.ylabel('Number of Employees')
         plt.show()

         data.left.value_counts()
```

Figure09: Employee Left Visualization

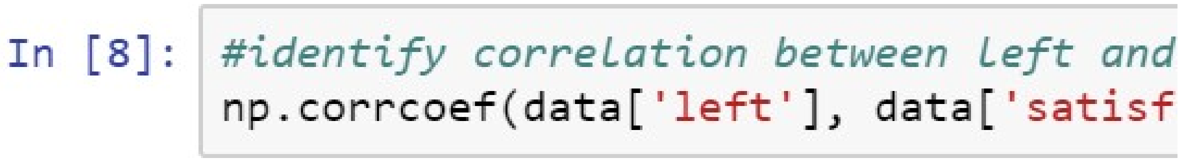
We can see that the satisfaction level is less for the employees who left the organization.



```
In [13]: satisfaction_lvl=data.groupby('satisfact
         plt.bar(satisfaction_lvl.index.values, sa
         plt.xlabel('Satisfaction Level')
         plt.ylabel('Number of Employees')
         plt.show()
```

Figure10: Satisfaction Level of Employees Left

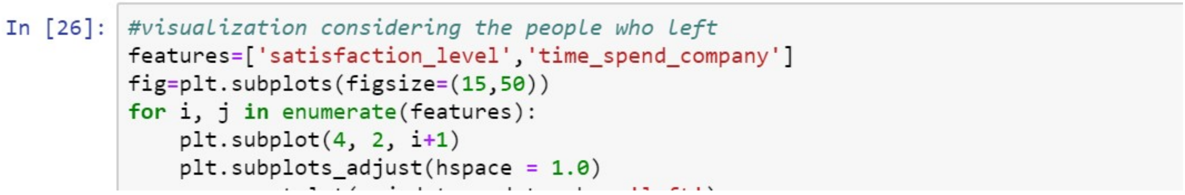
This is also verified from our correlation coefficient analysis, where correlation coefficient between **'left'** and **'satisfaction_level'** was negative. From the definition of correlation, it means as the value of '**satisfaction_level**' decreases, the value of '**left**' increases.



```
In [8]: #identify correlation between left and
        np.corrcoef(data['left'], data['satisf
```

Figure11: Correlation Coefficient Between Left and Satisfaction Level

Further visualization gives us more clarity on the effects of the selected features on the employees who left the organization. We can see that the people who left were less satisfied with the organization, and the employee with five-year experience is leaving more. For the feature time with company, the three-year mark looks like a time to be a crucial point in an employee's career. Most of them quit their job around the three-year mark. Another important point is 6-years point, where the employee is very unlikely to leave. The positive correlation coefficient between '**left**' and '**time_spend_company**' and between '**left**' and '**average_monthly_hours**' also gives us the idea that longer time spent at company and more work reduces the employee's chances of attrition.



```
In [26]: #visualization considering the people who left
         features=['satisfaction_level','time_spend_company']
         fig=plt.subplots(figsize=(15,50))
         for i, j in enumerate(features):
             plt.subplot(4, 2, i+1)
             plt.subplots_adjust(hspace = 1.0)
```

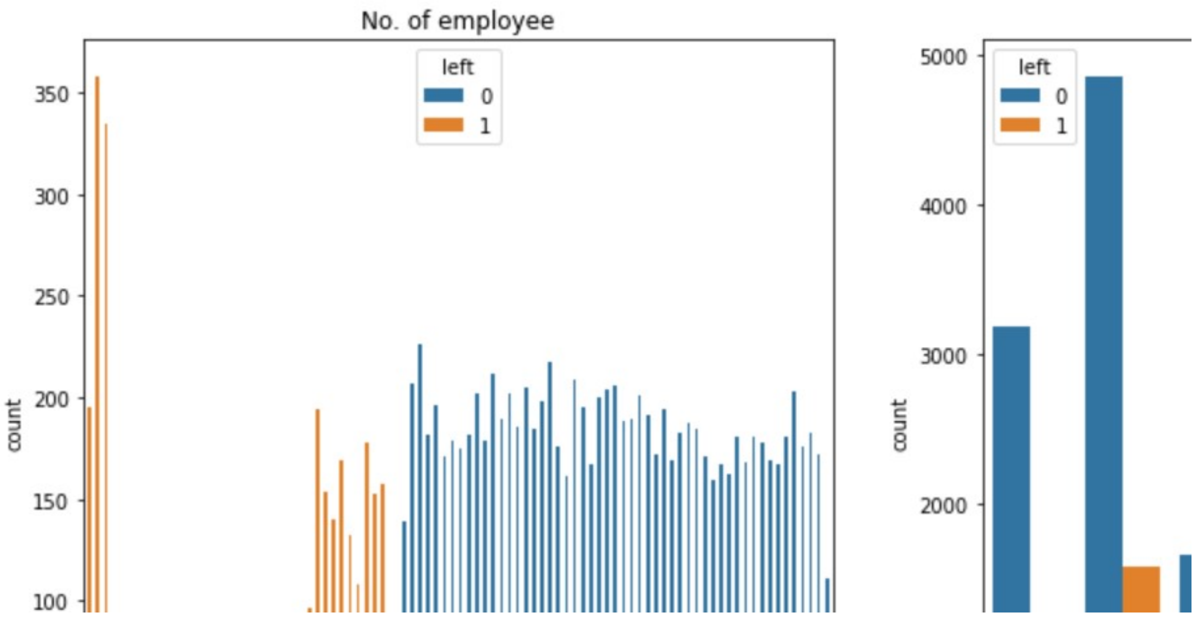


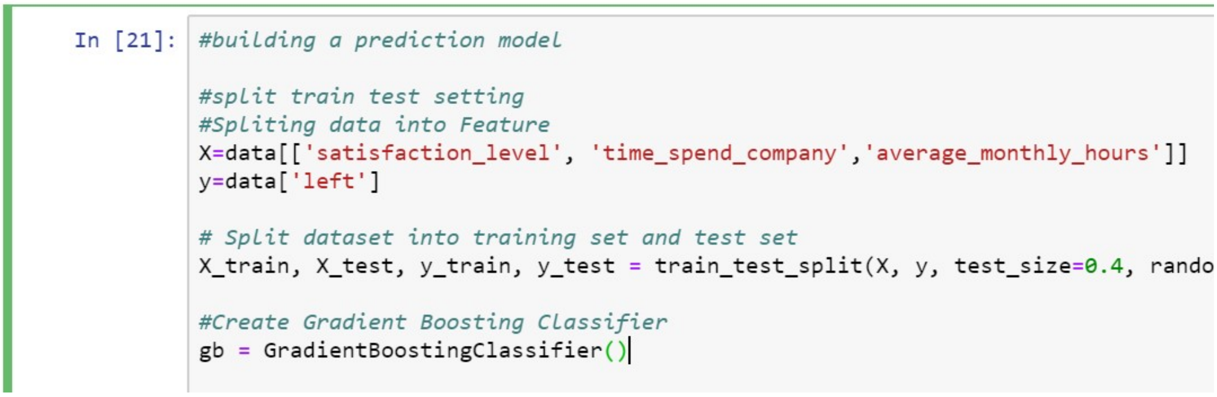
Figure12: Data Visualization (Cont.)

4.3 Predictive Model Building

The principal focus of the research was that whether we can identify the important features behind employee attrition and use them to build a predictive model that can help the organization management with attrition prediction.

What we are going to do here is to divide the dataset into a training set and a test set using function **train_test_split()**. It requires to pass 3 parameters - **features, target, and test_size**. We have also used **random_state** to select records randomly. The dataset is split into two sections in ratio of 60:40. It means 60% data will be used for model training and 40% will be used for model testing.

In the model we are going to predict employee attrition using Gradient Boosting Classifier. We created a Gradient Boosting classifier object using **GradientBoostingClassifier ()** function. After that, we fit our model on train set using **fit ()** and performed prediction on the test set using **predict ()**.

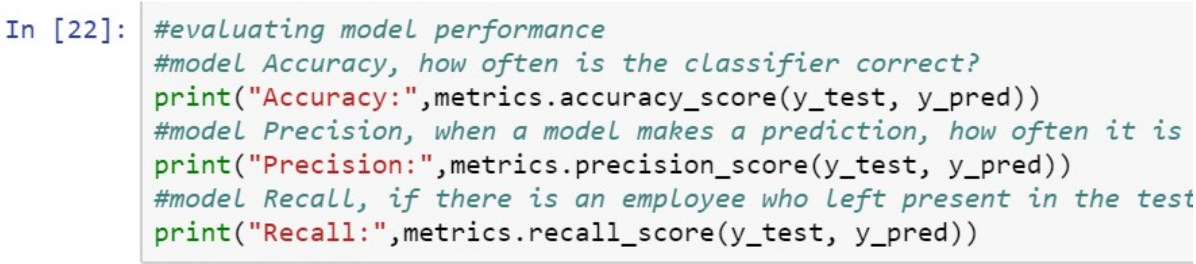


```
In [21]: #building a prediction model

         #split train test setting
         #Spliting data into Feature
         X=data[['satisfaction_level', 'time_spend_company','average_monthly_hours']]
         y=data['left']

         # Split dataset into training set and test set
         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4, rando

         #Create Gradient Boosting Classifier
         gb = GradientBoostingClassifier()
```

Figure13: Predictive Model Building

To evaluate the performance of the model, we used the widely accepted performance evaluation metrics from scikit-learn:

- accuracy (fraction of samples predicted correctly)
- precision (fraction of positive events predicted correctly)
- recall (fraction of positive events that are actually positive)



```
In [22]: #evaluating model performance
         #model Accuracy, how often is the classifier correct?
         print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
         #model Precision, when a model makes a prediction, how often it is
         print("Precision:",metrics.precision_score(y_test, y_pred))
         #model Recall, if there is an employee who left present in the test
         print("Recall:",metrics.recall_score(y_test, y_pred))
```

Figure14: Performance Evaluation

Accuracy: we can see our model’s classification rate of 96%.

Precision: How often the model is correct when it makes a prediction. In our case, when the Gradient Boosting model forecasted an employee is going to leave, that employee actually left 93% of the time.

Recall: If there is an employee who left is present in the test set, our model was able to identify it 91% of the time.

4.4 Summary

From our analysis, we found out that according to our dataset, employees less satisfied are prone to leave the organization. Employees who have been with the company longer and have involved with work are less likely to churn out. Our prediction model was able to show well enough performance to forecast an employee leaving the organization given values of the most important features.

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

5.1 Findings and Contributions

Organizations must pay more attention towards increasing the employees work related satisfaction if they want to retain their valuable assets. The more involved he is with work, and the time he has spent with the company are also very important in the employees decision on when to leave to organization, therefore it is very important for the organization management to focus on these two factors as well.

5.2 Limitations

There have been several limitations to be acknowledged, related to our research work. The dataset could have been larger with more employee attributes that could have even more significance in identifying the more important contributing factors. If we could get more data in our hand, the research work would have been much more efficient. The training data would be higher in number and as a result, the prediction outcome would have come much more accurate.

5.3 Recommendations for Future Works

Future research work on this topic may include collecting a larger dataset and include more important contributing factors and features. This will ensure the outcome of the research become more accurate and start to contribute properly in employee attrition forecasting for any organization.

REFERENCES

[1]. Abelson, M., B. Baysinger (1984), “Optimal and dysfunctional turnover: Toward an organizational level model,” Academy of Management Review, Vol. 9 No.2, pp. 331–341.

[2]. Arnold, H.J. and Feldman, D.C., (1982), “A multivariate analysis of the determinants of job turnover,” Journal of Applied Psychology, Vol. 67, No.3, pp. 350-360.

[3]. Arthur, W., Bell, S., Donerspike, D., &Villado, A.,(2006), “ The use of Person-Organization fit in employment decision making; An assessment of its criterion related validity,” Journal of Applied Psychology, Vol.91, pp. 786-801.

[4]. Barrick, M.R., & Zimmerman, R.D.,(2005), “Reducing voluntary turnover, avoidable turnover through selection,” Journal of Applied Psychology, Vol. 90, pp.159-166

[5]. Berg, T.R., (1991), The importance of equity perception and job satisfaction in predicting employee intent to stay at television stations. Group and Organization Studies, Vol.16, No.3, pp. 268-284.

[6]. Bliss WG (2007), “Cost of employee turnover”, available atwww.isquare.com/turnover.

[7]. Boswell, W.R., Boudreau, J.W., &Tichy, J., (2005), “The relationship between employee job change and job satisfaction: The honey moon-hangover effect,” Journal of Applied Psychology, Vol.47, pp.275-301.

[8]. Buckley (2004) “The attrition of both new and experienced teachers was a great challenge for schools and school administrators throughout the United States”, available at http://www.edfacilities.org/pubs/teacherretention.cfm

[9]. Cappelli, P. (2008), “Talent management for the twenty-first century”, Harvard Business Review, March, 74-81.

[10]. Chaminade B (2007), “A retention checklist: how do you rate?” available at www.humanresourcesmagazine.co.au.

[11]. Cotton, J.L. and Tuttle, J.F., (1986), “Employee turnover: A meta-analysis and review with implications for research,” Academy of Management Review, Vol.11, No.1, pp. 55-70.

[12]. Dickter, D.N., Roznowski, M. and Harrison, D.A., (1996), “Temporal tempering: An event history analysis of the process of voluntary turnover,” Journal of Applied Psychology, Vol.81, pp.707–716.

[13]. Gerhart, B., (1990)., “Voluntary turnover and alternative job opportunities,” Journal of Applied Psychology, Vol.75, No.5, pp. 467- 476.

[14]. Glebbeek, A.C., & Bax, E.H.,(2004), “ Is high employee turnover really harmful? An empirical test using company records,” Academy of Management Journal, Vol.47, pp. 277-286.

[15]. Hendricks S (2006), “Recruitment & retention of appropriately skilled people for the public service to meet the challenges of a developmental state”, Conference of senior managers of the Free State Provincial government, local authorities, state agencies & the business sector.

[16]. Hinkin, T.R., & Tracey, J.B.,(2000), “The cost of turnover: Putting a price on the learning curve,” Cornell Hotel & Restaurant Administration Quarterly, Vol 41, pp.14-21.

[17]. Oyelade, O. J , Application of k-Means Clustering algorithm for prediction of Students’ Academic Performance

https://arxiv.org/ftp/arxiv/papers/1002/1002.2425.pdf

[18]. Susmita Datta and Somnath Datta, “Comparisons and validation of statistical clustering techniques for microarray gene expression data,” Bioinformatics, vol. 19, pp.459–466, 2003.

[19]. Rousseeuw P. J, “A graphical aid to the interpretation and validation of cluster analysis,” Journal of Computational Appl Math, vol 20, pp. 53– 65, 1987.

[20]. Sharmir R. and Sharan R., “Algorithmic approaches to clustering gene expression data,” In current Topics in Computational Molecular Biology MIT Press; pp. 53-65, 2002.

[21]. Mucha H. J., “Adaptive cluster analysis, classification and multivarite graphics,”Weirstrass Institute for Applied Analysis and Stochastics, 1992.

[22]. Fahim A. M., Salem A. M., Torkey F. A. and Ramadan M. A., “An efficient enhanced k-means clustering algorithm,” Journal of Zhejiang University Science A., pp. 1626–1633, 2006